

BAO annotations of the DrugMatrix enable bioactivity-based target- relationship analyses and demonstrate incorporation of external datasets into BARD

Tudor Oprea, Stephan Schurer, Ahsan Mir, Uma
Vempati, Jeremy Yang, Oleg Ursu, Cristian Bologa

PI: Larry Sklar (NIH U54 MH084690)



Contributors & Stakeholders

Direct Contributors –MLPCN Centers (*generated ~ 85% of PubChem data*)



Stakeholders & Advisors: Architecture Advisory (Wash U), **Technical Advisory Group** (NIH, EBI, NIBR, Takeda), **Requirements & Usability Feedback Group**



Worcester Polytechnic Institute



* MLPCN Center



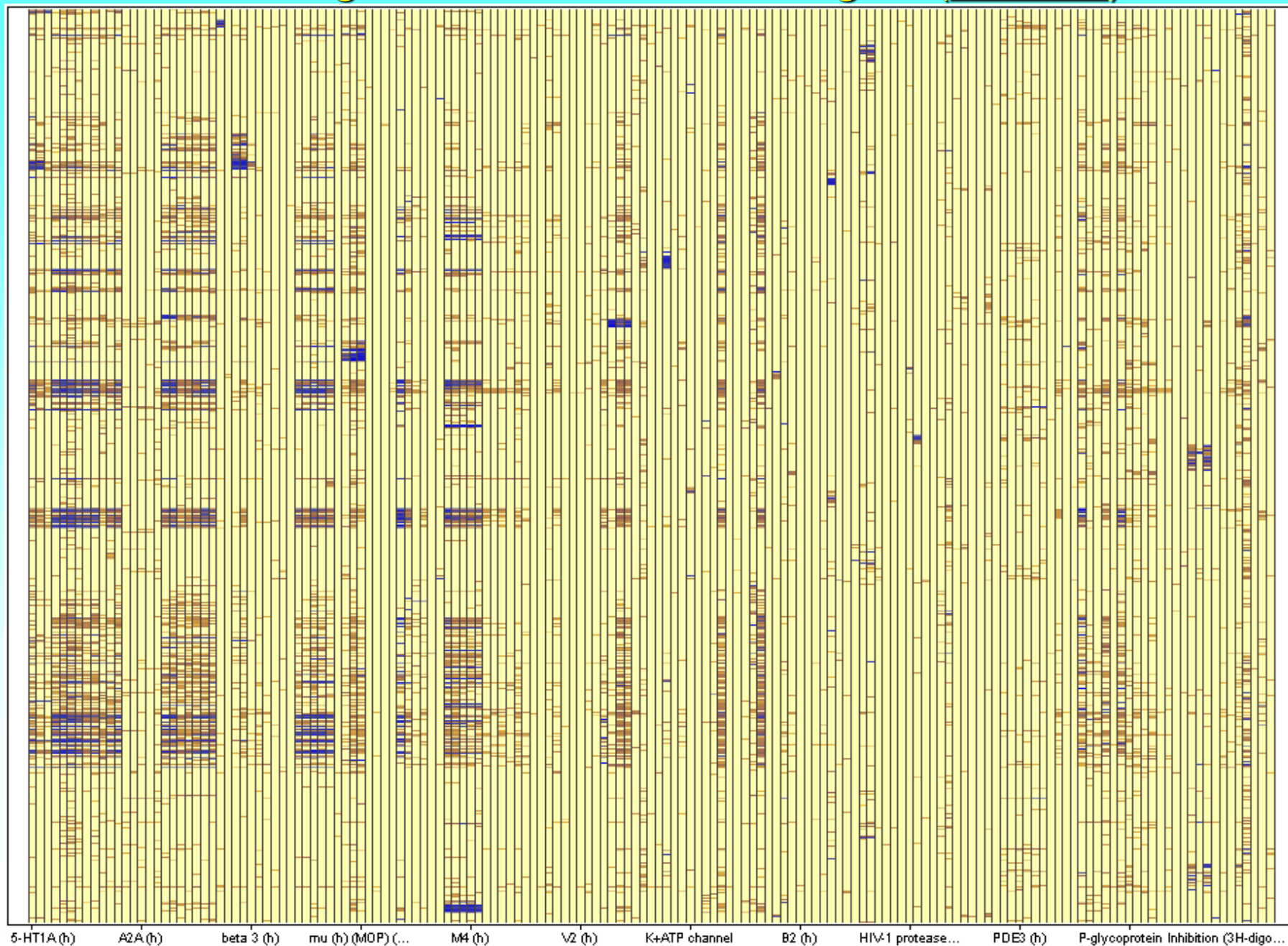
BioAssay Research Database



*BARD mission is to enable novice and expert scientists to effectively utilize MLP data to **generate and test new hypotheses***

- Unique collaboration amongst NIH and academic centers with expertise in **screening** and **software development**
- Developed as an **open-source**, industrial-strength platform to support **public** translational research.
- Provides opportunity to address existing cheminformatics barriers
 - **Deploy** predictive models
 - Foster new methods to **interpret** chemical biology data
 - Enable **private** data sharing and **collaborative** discovery
 - **Develop** and **adopt** an **Assay Data Standard** with tools to:
 - Annotate assays to a **minimum standards** and **definitions**
 - **Integrate** and **extend** existing ontologies for meaningful assay descriptions
 - **Enable** assay creation, registration and modification
 - **Provide** an easy-to-use portal and an advanced desktop client

737 Drugs Measured on 159 Targets (CEREP)

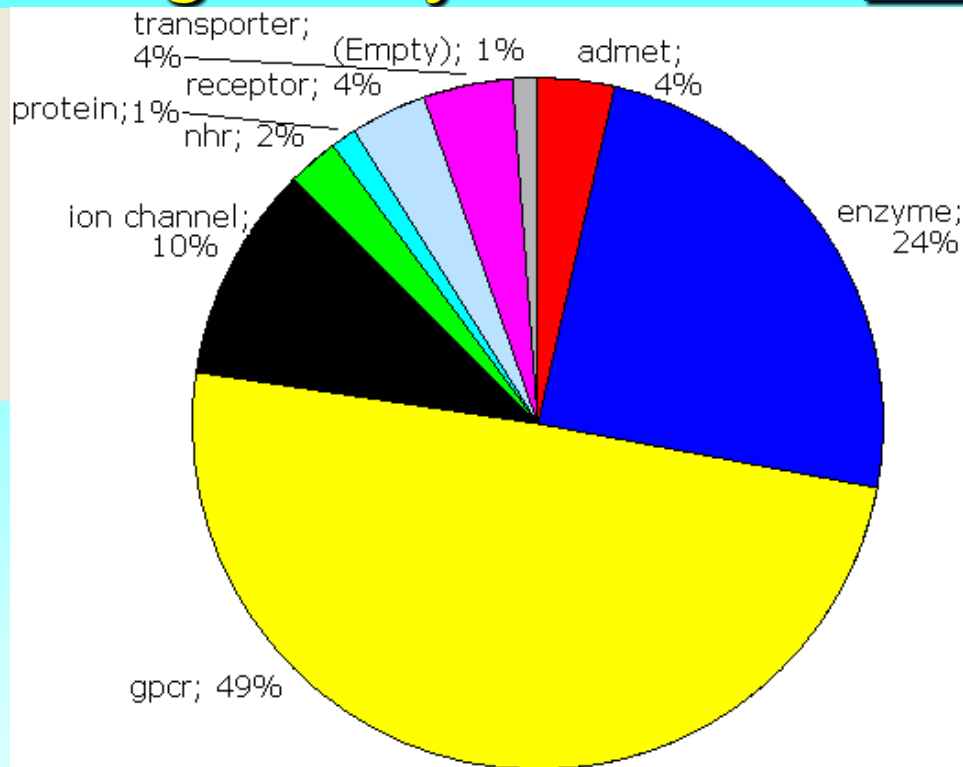


With Scott Boyer (AstraZeneca)

Targets by Class in CEREP

TargetClass

- ☒ admet
- ☒ enzyme
- ☒ gpcr
- ☒ ion channel
- ☒ nhr
- ☒ protein
- ☒ receptor
- ☒ transporter



2211 Compounds:

- 1225 Drugs;
- 45 veterinary drugs;
- 112 prodrugs;
- 14 active metabolites.

Of the 188 total targets, 20 (DMEs) were removed.

The less specific “ADMET” assays were kept for comparison


Target Class	Enzyme	GPCR_A	GPCR_B	GPCR_C	“ADMET”	Proteins	Ion Channels	NHRs	Receptors	Transporters
Nr Targets	41	57	24	2	6	2	18	4	7	7
Average Nr Actives	72	329	76	14	134	40	164	125	160	322
Nr Biased Targets	2	39	3	0	1	0	5	0	2	5
Average Nr Biased Actives	234	444	224	0	317	0	450	0	471	409


Total *potential* activities: **371,448**; Total *observed* activities: **31,264**; → *Probability*: 8.41%

Definition: “biased targets” as those who exceed the 8.41 probability in this dataset.

Biased targets account for 24,015 (76.81%) of the activities in the CEREP dataset!

DrugMatrix Dataset Origin

EMBL-EBI  Services Research Traini


 **ChEMBL**

Activity Source Filter

EBI > Databases > Small Molecules > ChEMBL Database > [Ligand Search](#)

Search ChEMBLdb... Compounds Targets Assays Documents [Activity Source Filter](#)

ChEMBLdb Ligand Search Target Search Browse Targets Browse Drugs Drug Approvals



C
N
O
S
F
Cl
Br
I
P
X

ChEMBLdb Statistics

- DB: ChEMBL_16
- Targets: 9,844
- Compound records: 1,487,579
- Distinct compounds: 1,295,510
- Activities: 11,420,351
- Publications: 50,095
- [Release Notes](#)

ChEMBL Blog

- [New Drug Approvals 2013 - Pt. XII - Technetium Tc 99m Tilmanocept](#)

List Search

☐ SMILES Search

Please enter keywords, or

Biologicals Blast

JME Molecular Editor (c) Peter Ertl

<https://www.ebi.ac.uk/chembl/db/>



DrugMatrix Dataset Origin

EMBL-EBI



ChEMBLdb

Malaria Data

ChEMBL-NT

Kinase SARf

GPCR SARf

DrugEBility

Downloads

Web Service

FAQ

ChEMBLdb

DB: ChEM

Targets: 9

Compound

1,487,579

Distinct co

1,295,510

Activities:

Publication

Release N

ChEMBL Bk

New Drug

2013 - Pt

Technetiu

Tilmanoce

Selected Bioactivity Sources

Selected	Source	Assay Counts	Activity Counts
<input checked="" type="checkbox"/>	Scientific Literature	703683	4044415 (35.41%)
<input checked="" type="checkbox"/>	TP-search Transporter Database	3592	6765 (0.06%)
<input checked="" type="checkbox"/>	PubChem BioAssays	2210	6571997 (57.55%)
<input checked="" type="checkbox"/>	Open TG-GATEs	1376	179525 (1.57%)
<input checked="" type="checkbox"/>	Millipore Kinase Screening	468	73944 (0.65%)
<input checked="" type="checkbox"/>	GSK Published Kinase Inhibitor Set	454	168717 (1.48%)
<input checked="" type="checkbox"/>	Sanger Institute Genomics of Drug Sensitivity in Cancer	352	5984 (0.05%)
<input checked="" type="checkbox"/>	Guide to Receptors and Channels	344	801 (0.01%)
<input checked="" type="checkbox"/>	DrugMatrix in vitro pharmacology assays	132	229944 (2.01%)
<input checked="" type="checkbox"/>	Drugs for Neglected Diseases Initiative (DNDi)	62	11554 (0.1%)
<input checked="" type="checkbox"/>	OSDD Malaria Screening	21	226 (0%)
<input checked="" type="checkbox"/>	WHO-TDR Malaria Scr	https://www.ebi.ac.uk/chembl/db/index.php/downloads#	
<input checked="" type="checkbox"/>	St Jude Malaria Screening	16	5456 (0.05%)

DrugMatrix @ NTP



National Toxicology Program
Department of Health and Human Services

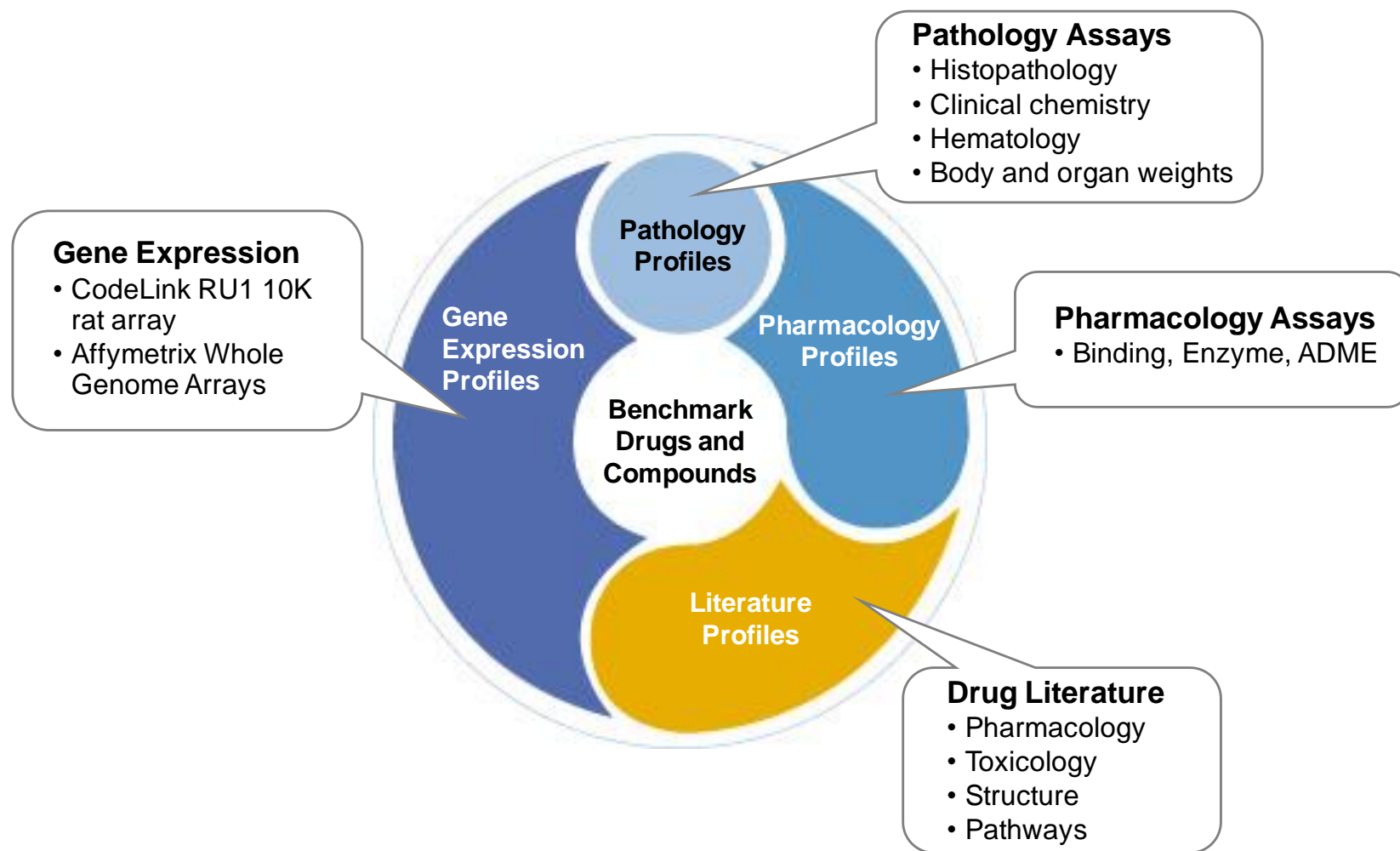
DrugMatrix®

DrugMatrix is the scientific communities' largest molecular toxicology reference database and informatics system. DrugMatrix is populated with the comprehensive results of thousands of highly controlled and standardized toxicological experiments in which rats or primary rat hepatocytes were systematically treated with therapeutic, industrial, and environmental chemicals at both non-toxic and toxic doses. Following administration of these compounds in vivo, comprehensive studies of the effects of these compounds were carried out at multiple time points and in multiple target organs. These studies included extensive pharmacology, clinical chemistry, hematology, histology, body and organ weights, and clinical observations. Additionally, a curation team extracted all relevant information on the compounds from the literature, the Physicians' Desk Reference, package inserts, and other relevant sources. The heart of the DrugMatrix database is large-scale gene expression data generated by extracting RNA from the toxicologically relevant organs and tissues and applying these RNAs to the GE Codelink™ 10,000 gene rat array and more recently the Affymetrix whole genome 230 2.0 rat GeneChip® array. DrugMatrix contains toxicogenomic profiles for 638 different compounds; these compounds include FDA approved drugs, drugs approved in Europe and Japan, withdrawn drugs, drugs in preclinical and clinical studies, biochemical standards, and industrial and environmental toxicants. Contained in the database are 148 scorable genomic signatures derived using MOSEK computational software that cover 96 distinct phenotypes. The signatures are informative of organ-specific pathology (e.g., hepatic steatosis) and mode of toxicological action (e.g., PXR activation in the liver). The phenotypes cover a number of common target tissues in toxicity testing (including liver, kidney, heart, bone marrow, spleen and skeletal muscle). The primary value that DrugMatrix provides to the toxicology community is in its capacity to use toxicogenomic data to perform rapid toxicological evaluations. Further value is provided by DrugMatrix ontologies that help characterize mechanisms of pharmacological/toxicological action and identify potential human toxicities. Overall, DrugMatrix allows a toxicologist to formulate a comprehensive picture of toxicity with greater efficiency than traditional methods.





DrugMatrix Database Content

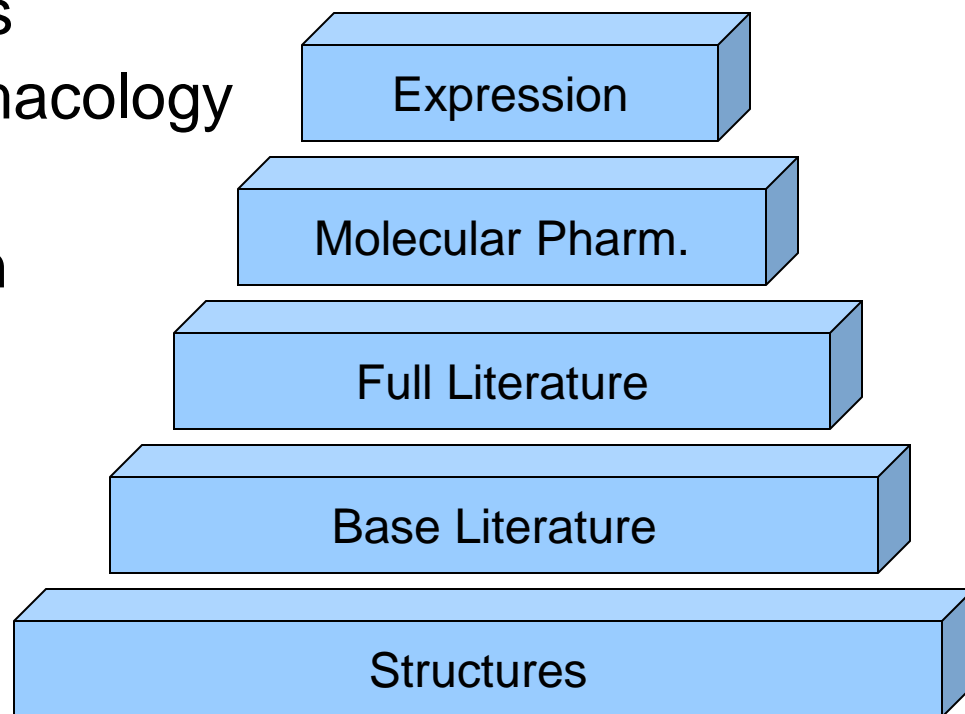




Progression of Content

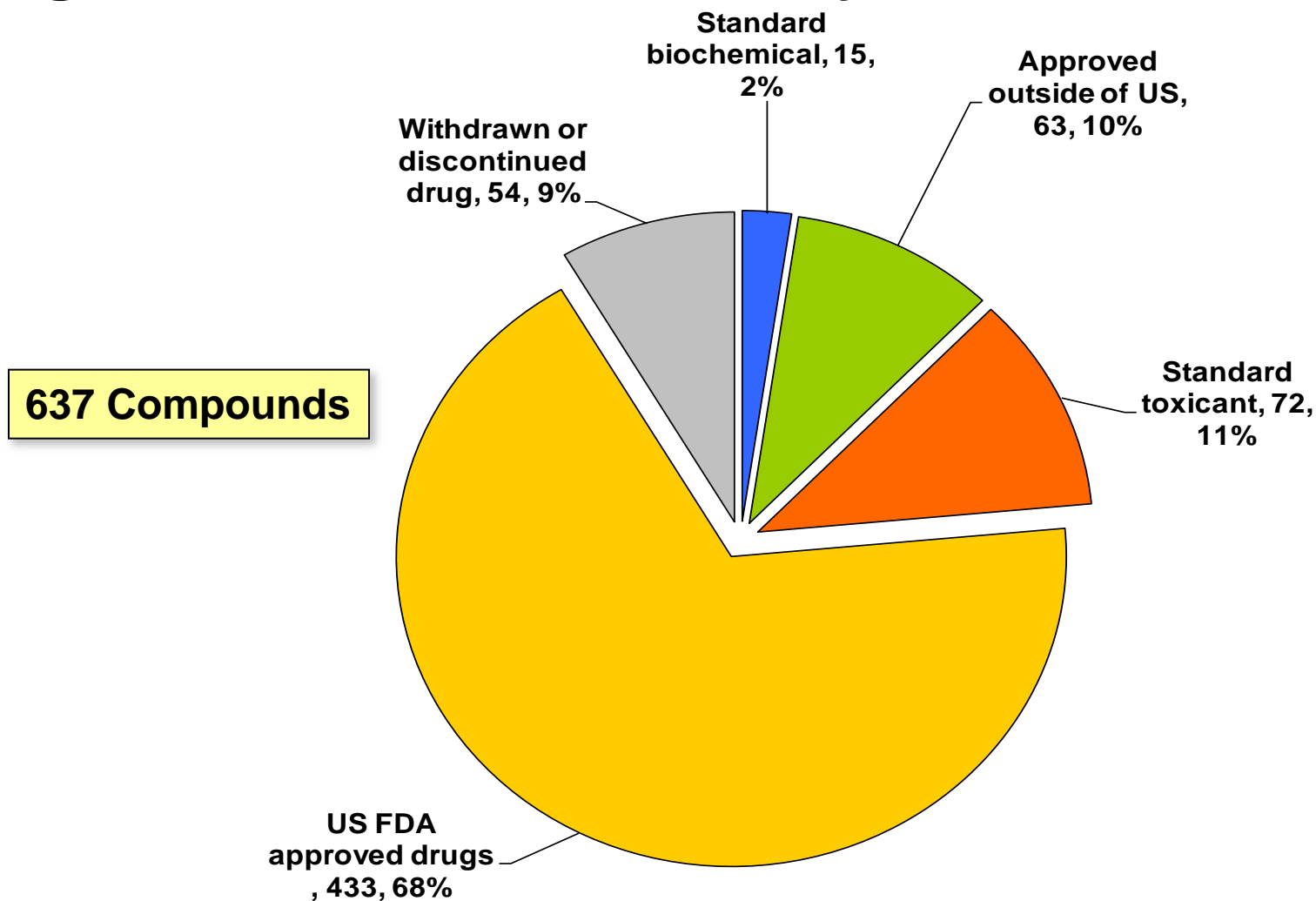
Numbers of Compounds:

- 637 in Expression Studies
- 867 with Molecular Pharmacology
- ~900 with Full Curation
- ~2000 with Base Curation
- ~8000 with Structures





DrugMatrix Chemical Diversity



More data – not always more knowledge

- Several public resources of screening results exist
 - PubChem
 - PDSP
 - ChEMBL
 - Binding DB
 - ...
- But there are many challenges
 - Syntactic, structural, semantic heterogeneity problems
 - Incomplete or no annotations
 - Lack of standardized metadata
 - Project context not well defined (assay relations)

BAO scope and purpose



- BAO to describe assays and screening results
 - Defines relevant assay and result annotations
 - Provides controlled terminology
 - Formalizes knowledge of assays and screening results
 - Described and formalizes screening campaigns
- BAO addresses data problems to facilitate
 - Leveraging existing data in discovery projects
 - Global analyses across diverse data sets
 - Integration data from different resources

BioAssay Ontology (BAO) history



- Started November 2009 NIH funded
- Version 0.9 released in September 2010
- BAO 2.0 released Aug 2013
- Systematic assay annotations (PubChem)
- BAOsearch: Semantic Web Software System:
<http://baosearch.ccs.miami.edu/>
(demonstration project; open source)
- Several BAO papers and presentations:
development, application, data analysis
- More info: <http://bioassayontology.org/>

BAO projects and impact



- Several projects / organizations using or considering BAO



- BARD
- EU OpenScreen
- OpenPHACTS (Astra Zeneca)
- ChEMBL
- LIFE (NIH LINCS program)
- More: NCBI Bioportal (and coming)



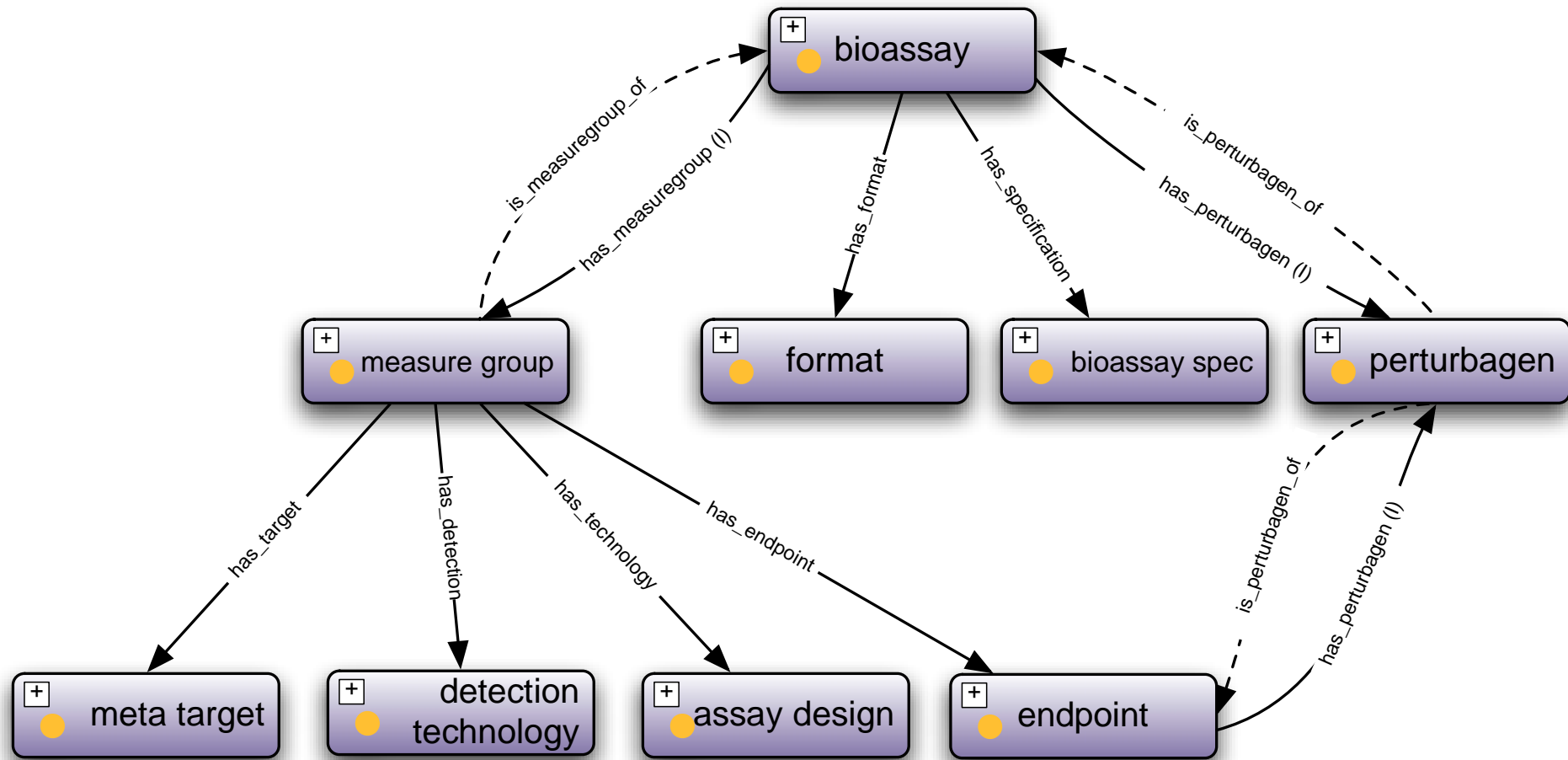
- Related funded research projects

- BARD: <http://bard.nih.gov/>
- LINCS Information FramEwork (LIFE): <http://lifekb.org>
- Regenbase: <http://www.regenbase.org/>

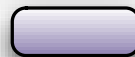
So what do we mean by ontology?

- Controlled vocabularies / thesauri: describe what things mean (link terms to definitions)
 - Share knowledge in a common language
 - Organize knowledge
- Formalization of knowledge using logical axioms
- Explicit specification (OWL-DL)
 - Building models (abstracting a knowledge domain)
 - Computing with knowledge (reasoning machines)
 - Exchanging information (reconcile knowledge on a global scale)

BAO original outline to describe assays



Legend



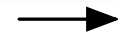
Ontology class



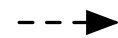
More subclasses



Primitive class

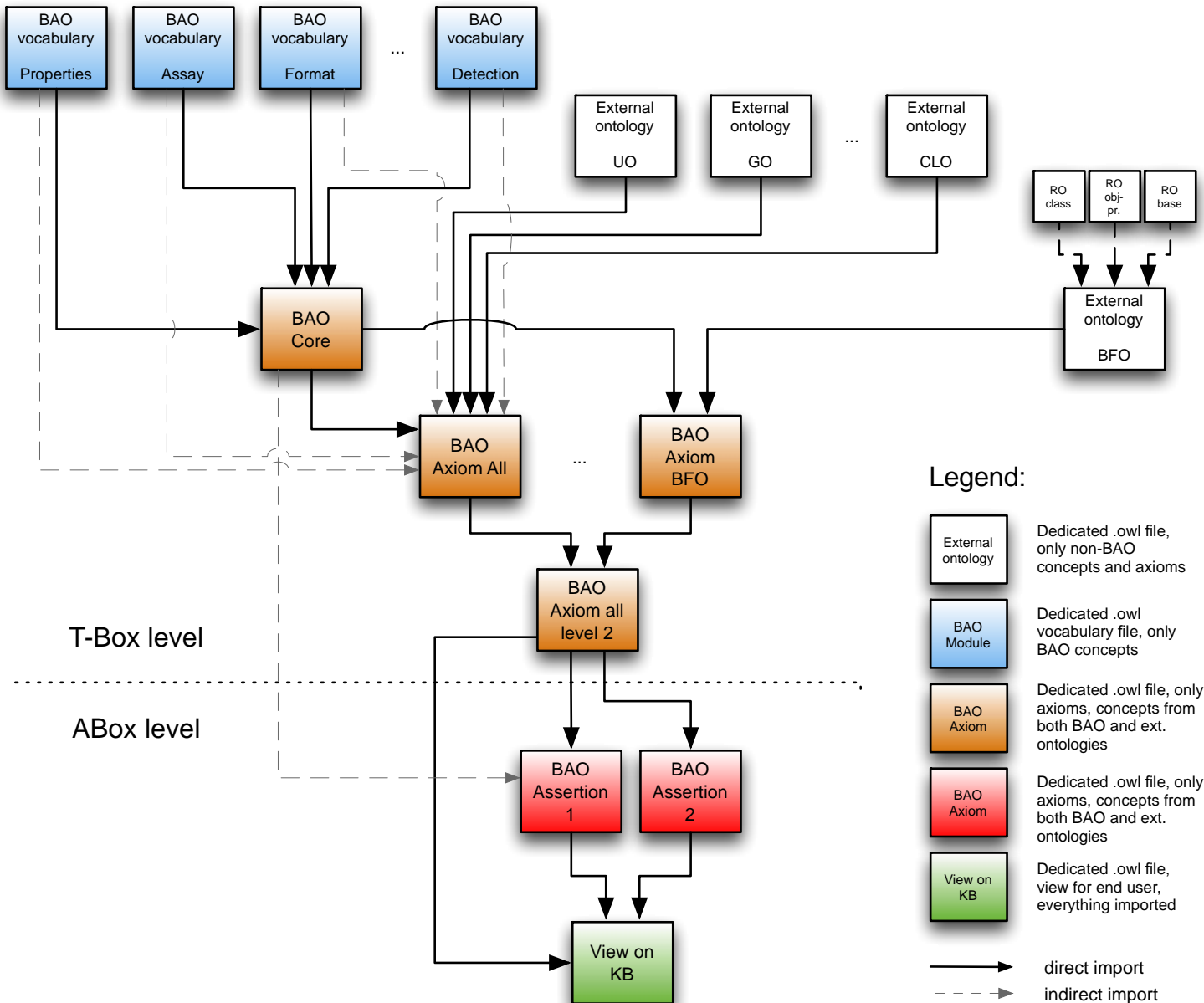


Asserted relation, (I) is inverse

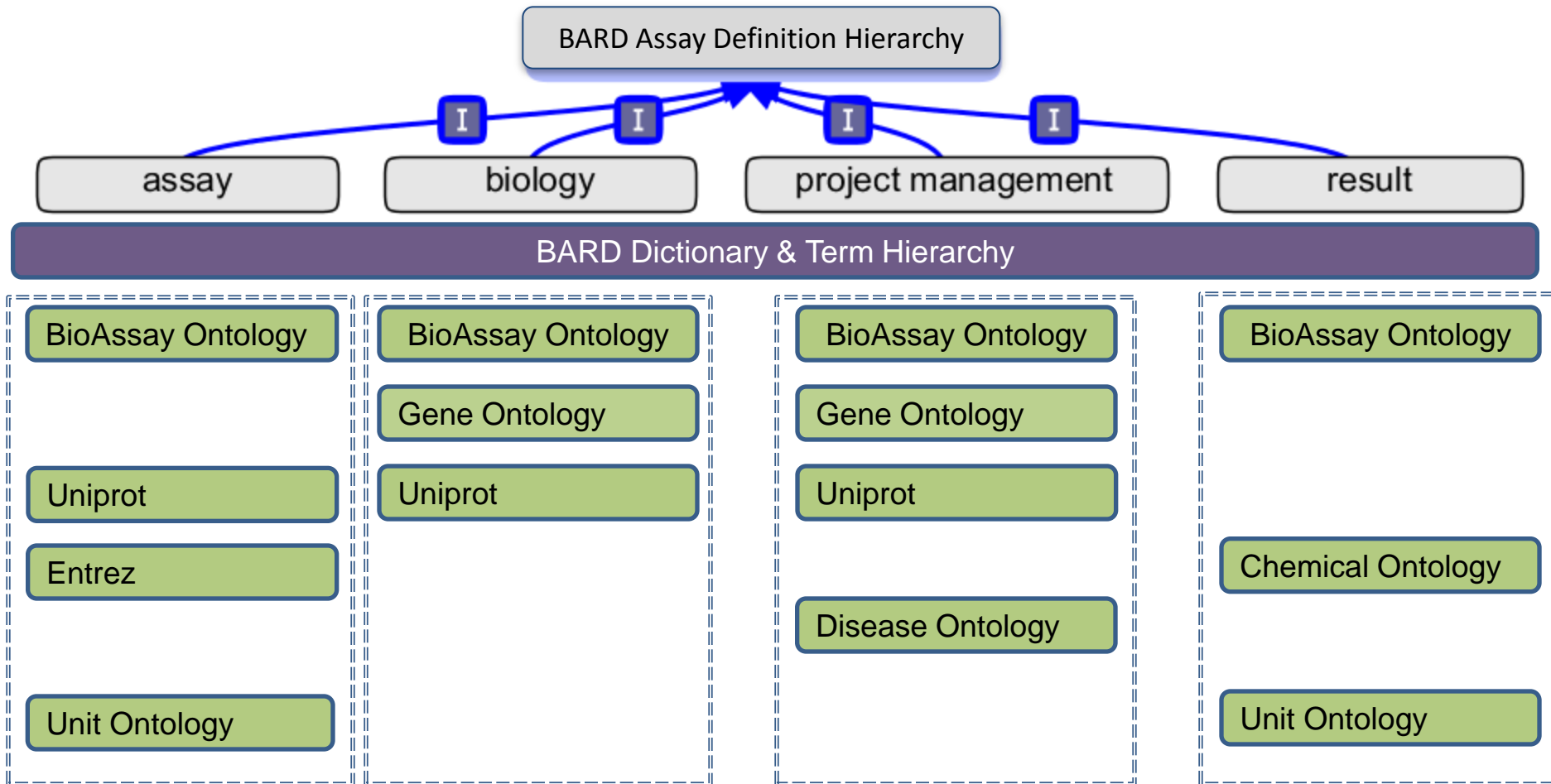


Inferred relation

BAO 2.0 Modularization



BARD Data Dictionary & Associated Ontologies



BioAssay Research Database



*BARD mission is to enable novice and expert scientists to effectively utilize MLP data to **generate and test new hypotheses***

- Unique collaboration amongst NIH and academic centers with expertise in **screening** and **software development**
- Developed as an **open-source**, industrial-strength platform to support **public** translational research.
- Provides opportunity to address existing cheminformatics barriers
 - **Deploy** predictive models
 - Foster new methods to **interpret** chemical biology data
 - Enable **private** data sharing and **collaborative** discovery
 - **Develop** and **adopt** an **Assay Data Standard** with tools to:
 - Annotate assays to a **minimum standards** and **definitions**
 - **Integrate** and **extend** existing ontologies for meaningful assay descriptions
 - **Enable** assay creation, registration and modification
 - **Provide** an easy-to-use portal and an advanced desktop client



DrugMatrix subset curation

- Assay descriptions were curated manually, types of information
 - Assay type, assay method, detection method
 - Assay format, target protein, target species, protein preparation
 - Substrate, reference compounds and activity
- Curation method / ontology
 - BAO2.0 development
 - Excel annotation template (capturing part of the hierarchy)

871 Chemicals Measured in 131 Assays



IC50 data for 131 assays; greyed out columns are targets with zero hits; *IC50 data only*

DrugMatrix Promiscuity



Target Class	Enzyme	GPCR_A	GPCR_B	GPCR_C	"ADMET"	Proteins	Ion Channels	NHRs	Receptors	Transporters
Nr Targets	39	51	4	0	7	0	17	5	4	4
Nr Targets	41	57	24	2	6	2	18	4	7	7
Nr Biased Targets	7	25	0	0	5 p450s	0	5	3	2	4
Nr Biased Targets	2	39	3	0	1	0	5	0	2	5

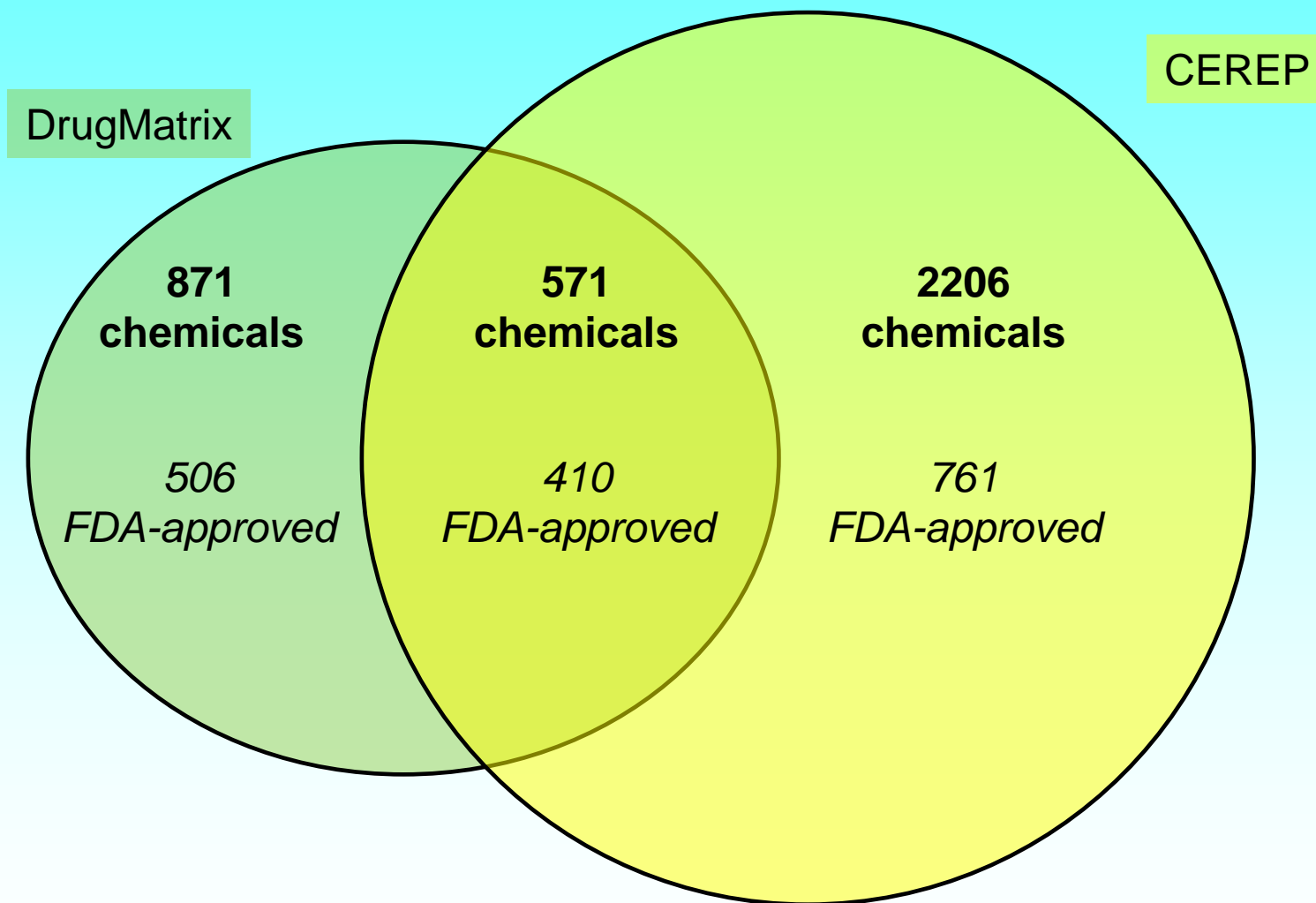
- The overall probability of IC50 bioactivity in the DM dataset is 3.45% (3939 out of 114101 cells)
- There are 62 chemicals that hit targets with higher probability. These include amitriptyline, chlorpromazine, olanzapine and thioridazine
- There are 51 targets that show bioactivity with probability higher than 3.45%. These include several GPCRs (adrenoceptors, 5HTRs, muscarinic, etc.), calcium channels, Cyp-P450s, 15LO, COX-1, COX-2, and kinases.

Some Observations

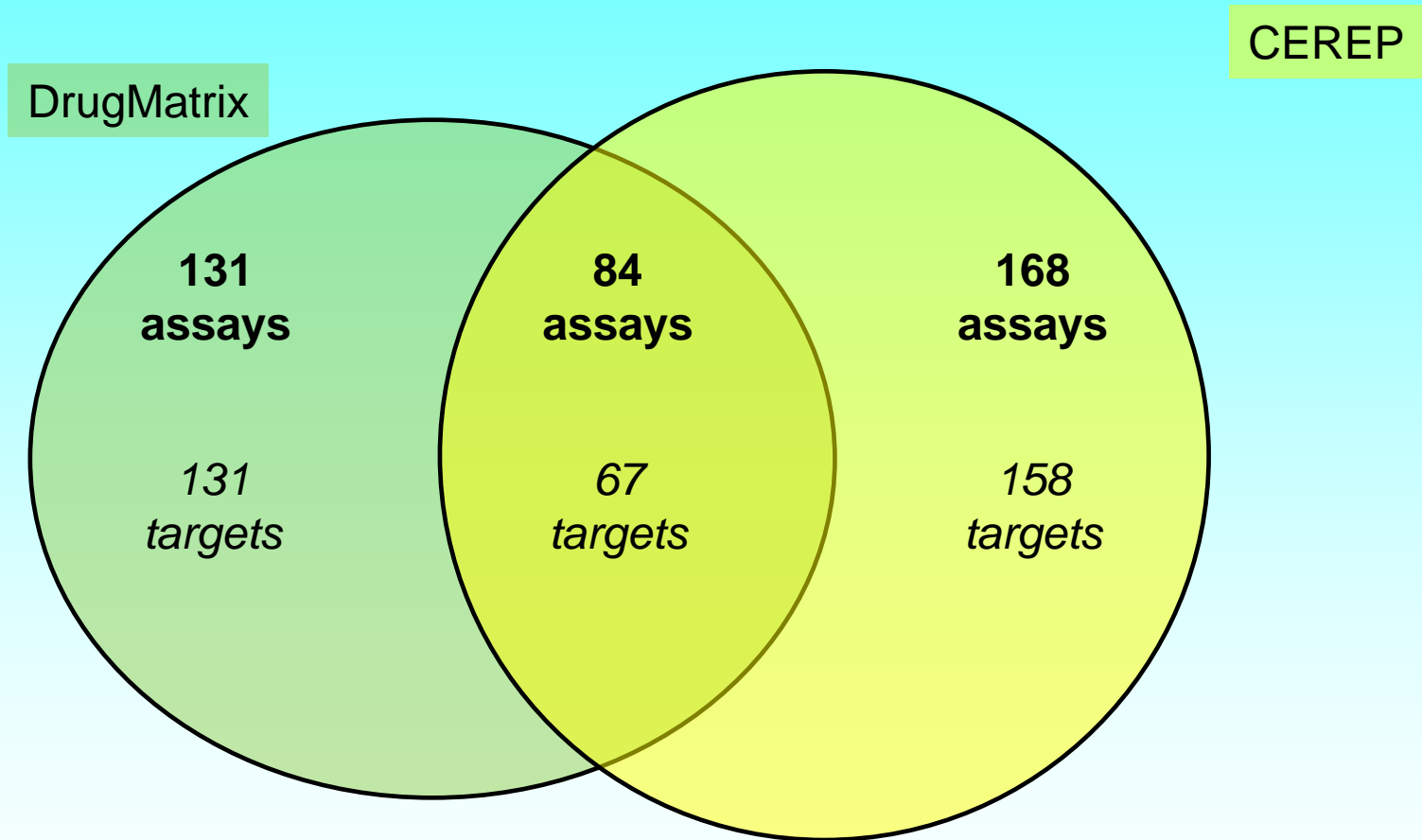


- Overall, 37 out of 131 targets needed manual curation at the target level; all needed ontology curation.
- Some mis-haps: “Sigma2 receptor” is a progesterone associated membrane protein; Imidazoline I2 receptor is an allosteric site of MAO.
- None of the links provided at the NTP / DM website are valid anymore (“ricerca” is now “eurofins”)
- As assays “evolve” vendors retire the information; difficult to track original assay information (we did capture it)

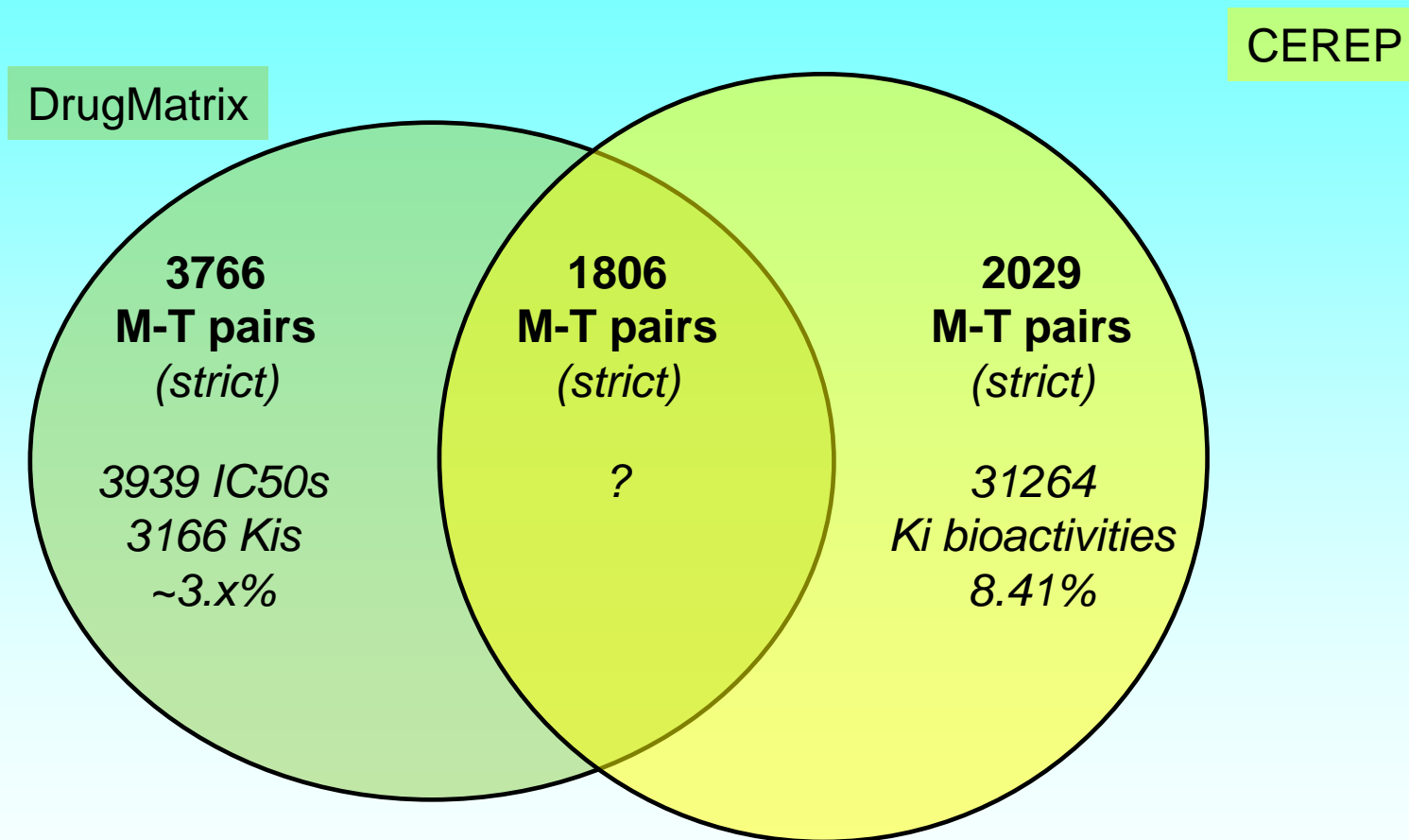
CEREP vs. DrugMatrix: Chemicals



CEREP vs. DrugMatrix: Targets



CEREP vs. DrugMatrix: Molecule-Target Pairs



To match bioactivities, assay conditions need to be evaluated as “similar” – that work could take months

Some Observations: 2



- Some assays have specificity of 70% - do we trust that?
- There are 4 assays where the KM is above 2000 uM, the reference inhibitor is below 10 uM – do we trust that?
- Several targets correlate at the “bioactivity fingerprint” level, on this particular set: e.g., adrenergics, calcium channels, muscarinics, share similar profile. A few surprises: D3R and 5HT6R; Imidazoline I2 and alpha-Ars
- *There is a similar situation in the CEREP Bioprint dataset.*

Instead of conclusions



- DrugMatrix was not designed by chemists (they screen polymers; Zn and Zr salts; at least 3 cpds with untreatable chemical structure).
- Data-by-data comparison with CEREP (another full matrix) reveals several molecular-target sets for which the overlap of bioactives is zero
- If you have an assay, you have information
- If you have two assays on the same target, you may have CONFUSION
- Where does knowledge come from?!
- What do we *really* know?
- *More curation is needed before the decision-making process is served well by experimental data*

Acknowledgments



- Stephan Schurer contributed slides and ontology terms
- Ahsan Mir & Uma Vempati assisted with curation
- Oleg Ursu, Cristian Bologna and Jeremy Yang assisted with data analytics
- *Scott Boyer provided the CEREP Bioprint dataset.*